Topological identification in networks of dynamical systems

Donatello Materassi* and Giacomo Innocenti**

* Department of Electrical and Computer Engineering,
University of Minnesota,
200 Union St SE, 55455, Minneapolis (MN)
mater013@umn.edu

** Dipartimento di Sistemi e Informatica,
Università di Firenze,
via di S. Marta 3, 50139 Firenze, Italy
giacomo.innocenti@gmail.com

***Abstract*—The paper deals with the problem of reconstructing the topological structure of a network of dynamical systems. A distance function is defined in order to evaluate the "closeness" of two processes and a few useful mathematical properties are derived. Theoretical results to guarantee the correctness of the identification procedure for networked linear systems with tree topology are provided as well. Finally, the application of the techniques to the analysis of an actual complex network, i.e. to high frequency time series of the stock market, is illustrated.**

I. Introduction

Under the influence of improved numerical tools, a significant interest for complex systems has been shown in many scientific fields. In particular, attention has been focused on networks, highlighting the emergence of complicated phenomena from the connection of simple models. To this regard, a relevant impulse has been provided by the advances in neural network theory, that has contributed to underline the importance of the connection topology in the realization of complex dynamics [1]. As a consequence, graph theory [2] has been successfully exploited to perform novel modeling approaches in several fields, such as Economics (see e.g. [3], [4], [5]), Biology (see e.g. [6], [7]) and Ecology (see e.g. [8], [9], [10]), especially when the investigated phenomena were characterized by spatial distribution and a multivariate analysis technique is preferred [11], [12].

To the best knowledge of the authors, there are very few theoretical results about the reconstruction of an unknown topology from data. In this paper, we will focus our attention on tree topology networks. Though its reduced complexity with respect to cyclic link structures, the tree connection model turns out to be particularly suitable to represent a large variety of processes. In particular, the tree network scheme results effective in the description of systems with transportation, such as water and power supply, air and rail traffic, vascular systems of living organisms and channel and drainage networks (see e.g. [13], [9], [14], [15], [16]). It is worth to highlight that this kind of models is deeply related to the idea of delay, that characterizes the connections as transportation media. It is also important to recall that in linear dynamical system theory the transfer function is a powerful representation tool for delayed processes [17], [18]. In many situations, when the topology to be reconstructed is a tree, the only observable nodes are the leaves. Then, the usual theoretical framework is almost always set in standard graph theory as in the Unweighted Pair Group Method with Arithmetic mean (UPGMA) [19]. Its application is mainly in the reconstruction of evolutionary trees, but it has been widely employed also in many other areas: communication systems and resource allocations. Theoretically, such a technique guarantees an exact reconstruction of a tree topology only on the strong assumption that an ultrametric is defined among the considered nodes. An approach based on system theory and identification tools is completely missing. Specifically, there are no approaches considering explicitly the possibility of dynamics among the nodes. While dynamical networks have been deeply studied and analyzed in automatic control theory, the question of reconstructing an unknown network of dynamical systems has not been formally investigated. In fact, in most applicative scenarios the network is given or it is the very objective of design. However, there are also some interesting situations where the network links are actually unknown, such as in biological neural networks, biochemical metabolic pathways and financial markets. Even though an acyclical topology may seem a quite reductive choice, given an intricate and connected topology, we may be interested into "approximating" it with a tree. Such an approximation could be considered "satisfactory" if the most important connections were captured.

In this manuscript we will develop a rigorous mathematical method to exactly identify the connections scheme of a tree topology network of noisy linear dynamical systems, providing a theoretical background for linear network modeling. In particular, in Section II we will introduce definitions and preliminary results which are useful to characterize the mathematical framework. In Section III our approach to topology reconstruction will be presented and sufficient conditions for an exact identification will be reported as well. In Section IV the theoretical results will be confirmed by practical implementations of the proposed technique, illustrated by means of numerical examples. In Section V, we will show that the identification of a tree

topology can provide useful information even for complex network. To this end, we will apply our technique to the analysis of high frequency real data originated by a portfolio of financial stocks. Some final conclusions in Section VI will end the manuscript.

Notation:
$E[\cdot]$: mean operator;
$R_{XY}(\tau) \doteq E[X(t)Y(t+\tau)]$: cross-covariance function of stationary processes;
$R_X(\tau) \doteq R_{XX}(\tau)$: autocovariance;
$\rho_{XY} \doteq \frac{R_{XY}}{\sqrt{R_X R_Y}}$: correlation index;
$\mathcal{Z}(\cdot)$: Zeta-transform of a signal;
$\Phi_{XY}(z) \doteq \mathcal{Z}(R_{XY}(\tau))$: cross-power spectral density;
$\Phi_X(z) \doteq \Phi_{XX}(z)$: power spectral density;
with abuse of notation, $\Phi_X(\omega) = \Phi_X(e^{i\omega})$;
$\lceil\cdot\rceil$ and $\lfloor\cdot\rfloor$: ceiling and floor function respectively;
$(\cdot)^*$: complex conjugate.

II. Problem set up

In this section we formally introduce a model to address noisy linear dynamical systems interconnected to form a tree topology and we also provide a quantitative tool to characterize the mutual dependencies.

Let us consider a network of n time-discrete SISO linear dynamical systems affected by additive noises. Then, let $H_j(z)$ be the transfer function of the j-th system, $\{X_j(k)\}_{k\in\mathbb{Z}}$ and $\{U_j(k)\}_{k\in\mathbb{Z}}$ its output and input signals respectively and $\{\varrho_j(k)\}_{k\in\mathbb{Z}}$ a zero-mean wide-sense stationary noise. Hence, each system can be represented according to the model:

$$X_j(k) = H_j(z)U_j(k) + \varrho_j(k) \quad \forall j = 1,\dots,n\,. \tag{1}$$

We stress that no assumptions on the causality of $H_j(z)$ have been done. Moreover, let the property

$$E[\varrho_j \varrho_i] = 0 \quad \forall j \neq i\,, \tag{2}$$

holds. Then, suppose that the input signal U_i of each node results the output of another process and that the systems of the network are connected to form a tree topology, preventing the presence of cycles.
In this paper we will formally address this kind of network according to the following definition.

Definition 1: Consider the ensemble of a rooted tree topology of n nodes N_j and a corresponding set of n linear time-discrete SISO systems affected by noise, described according to the model (1). Namely, assume N_i as the root node. Moreover, let $\{\varrho_j\}_{j=1,\dots,n}$ be zero-mean wide-sense stationary random processes satisfying (2), i.e. mutually not correlated zero-mean noises. Then, we define *Linear Cascade Model Tree (LCMT)* a dynamical network defined by the equation system

$$\begin{cases} X_1 = H_1(z)X_{\pi_1} + \varrho_1 \\ \dots \\ X_n = H_n(z)X_{\pi_n} + \varrho_n\,, \end{cases} \tag{3}$$

where $H_i(z) \equiv 0$ and the set $\{\pi_1,\dots,\pi_n\}$ is a permutation of $\{1,\dots,n\}$.

Definition 2: A LCMT is *well-posed* if $\Phi_{\varrho_j}(\omega) > 0$ for all ϱ_j, and for all ω

Assuming to have a complete statistical knowledge of each process $\{X_i\}_{i=1,\dots,n}$, we are interested in the identification of the links, which describe the tree characterizing the network topology. To this aim, hereafter we introduce some preliminary results, which can be exploited to define a mathematical tool for the quantitative characterization of the connections.

Let us consider two stochastic processes X_i, X_j and let $W_{ji}(z)$ be a time-discrete SISO transfer function. Hence, consider the quadratic cost

$$E\left[(\varepsilon_Q)^2\right]\,, \tag{4}$$

where

$$\varepsilon_Q \doteq Q(z)(X_j - W_{ji}(z)X_i)$$

and $Q(z)$ is an arbitrary stable and causally invertible time-discrete transfer function weighting the error

$$e_{ji} \doteq X_j - W_{ji}(z)X_i\,.$$

Then, the computation of the transfer function $\hat{W}(z)$ that minimizes the quadratic cost (4) is a well-known problem in scientific literature and its solution is referred to as the Wiener filter [18].

Proposition 3 (Wiener filter): The Wiener filter modeling X_j by X_i is the linear stable filter $\hat{W}_{ji}$ minimizing the filtered quantity (4). Its expression is given by

$$\hat{W}_{ji}(z) = \frac{\Phi_{X_iX_j}(z)}{\Phi_{X_i}(z)} \tag{5}$$

and it does not depend upon $Q(z)$. Moreover, the minimized cost is equal to

$$\min E\left[\varepsilon_Q^2\right] = \\ = \frac{1}{2\pi}\int_{-\pi}^{\pi} |Q(\omega)|^2 \left(\Phi_{X_j}(\omega) - |\Phi_{X_jX_i}(\omega)|^2 \Phi_{X_i}^{-1}(\omega)\right) d\omega\,,$$

and the corresponding error

$$\hat{e}_{ji} \doteq X_j - \hat{W}_{ji}(z)X_i$$

is not correlated with X_i, i.e.

$$E[\hat{e}_{ji}X_i] = 0\,. \tag{6}$$

Proof: See, for example, [17], [18]. ■

Since the weighting function $Q(z)$ does not affect the Wiener filter, but only the energy of the filtered error, we can choose $Q(z)$ equal to $F_j(z)$, the inverse of the spectral factor of $\Phi_{X_j}(z)$, that is

$$\Phi_{X_j}(z) = F_j^{-1}(z)(F_j^{-1}(z))^*\,. \tag{7}$$

In particular, it is worth recalling that $F_j(z)$ is stable and causally invertible [20]. Therefore, the minimum of cost (4) assumes the value

$$\min E[\varepsilon_{F_j}^2] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \left(1 - \frac{|\Phi_{X_jX_i}(\omega)|^2}{\Phi_{X_i}(\omega)\Phi_{X_j}(\omega)}\right) d\omega\,. \tag{8}$$

Observe that, due to such choice of $Q(z)$, the cost turns out to explicitly depend on the *coherence function* of the two processes:

$$C_{X_iX_j}(\omega) \doteq \frac{|\Phi_{X_jX_i}(\omega)|^2}{\Phi_{X_i}(\omega)\Phi_{X_j}(\omega)}. \tag{9}$$

Let us recall that the coherence function is not negative and symmetric with respect to ω. Moreover, it is also well-known that the cross-spectral density satisfies the Schwartz inequality and, thus, the coherence function results limited between 0 and 1. Therefore, according to the previous results, the cost (8) turns out to be dimensionless and not depending on the "energy" of the stochastic processes X_i and X_j.

The following result holds.

Proposition 4: In a well-posed LCMT, the binary function

$$d(X_i, X_j) \doteq \left[\frac{1}{2\pi}\int_{-\pi}^{\pi}\left(1 - C_{X_iX_j}(\omega)\right)d\omega\right]^{1/2} \tag{10}$$

is a metric.

Proof: The only non trivial property to be proved is the triangle inequality. Let $\hat{W}_{ji}(z)$ be the Wiener filter between X_i, X_j computed according to (5) and e_{ji} the relative error. The following relations hold:

$$\begin{aligned} X_3 &= \hat{W}_{31}(z)X_1 + e_{31} \\ X_3 &= \hat{W}_{32}(z)X_2 + e_{32} \\ X_2 &= \hat{W}_{21}(z)X_1 + e_{21}. \end{aligned}$$

Since $\hat{W}_{31}(z)$ is the Wiener filter between the two processes X_1 and X_3, it performs better at any frequency than any other linear filter, such as $\hat{W}_{32}(z)\hat{W}_{21}(z)$. So we have

$$\begin{aligned} \Phi_{e_{31}}(\omega) &\leq \Phi_{e_{32}}(\omega) + |\hat{W}_{32}(\omega)|^2\Phi_{e_{21}}(\omega) + \\ &+ \Phi_{e_{32}e_{21}}(\omega)\hat{W}^*_{32}(\omega) + \hat{W}_{32}(\omega)\Phi_{e_{21}e_{32}}(\omega) \leq \\ &\leq (\sqrt{\Phi_{e_{32}}(\omega)} + |\hat{W}_{32}(\omega)|\sqrt{\Phi_{e_{21}}(\omega)})^2 \quad \forall\, \omega \in \mathbb{R}. \end{aligned}$$

For the sake of simplicity we neglect to explicitly write the argument ω in the following passages. Normalizing with respect to Φ_{X_3}, we find

$$\frac{\Phi_{e_{31}}}{\Phi_{X_3}} \leq \frac{1}{\Phi_{X_3}}(\sqrt{\Phi_{e_{32}}} + |\hat{W}_{32}|\sqrt{\Phi_{e_{21}}})^2$$

and considering the 2-norm properties

$$\begin{aligned} &\left(\int_{-\pi}^{\pi}\frac{\Phi_{e_{31}}}{\Phi_{X_3}}d\omega\right)^{\frac{1}{2}} \leq \\ &\leq \left(\int_{-\pi}^{\pi}\frac{\Phi_{e_{32}}}{\Phi_{X_3}}d\omega\right)^{\frac{1}{2}} + \left(\int_{-\pi}^{\pi}\frac{|\Phi_{X_3X_2}|^2}{\Phi_{X_3}\Phi_{X_2}}\frac{\Phi_{e_{21}}}{\Phi_{X_2}}d\omega\right)^{\frac{1}{2}}, \end{aligned}$$

where we have substituted the expression of $\hat{W}_{32}$. Finally, observing that

$$0 \leq \frac{|\Phi_{X_3X_2}|^2}{\Phi_{X_3}\Phi_{X_2}} \leq 1,$$

we find

$$d(X_1, X_3) \leq d(X_1, X_2) + d(X_2, X_3).$$

■

III. Main result

In this section we exploit the coherence-based distance (10) to derive sufficient conditions to guarantee the exact reconstruction of the topology of a dynamical network. To this end, we first need to introduce a few definitions and technical lemmas.

Definition 5: We define "path" from N_i to N_j a finite sequence of $l > 0$ nodes $N_{\pi_1}, ..., N_{\pi_l}$ such that

- $N_{\pi_1} = N_i$
- $N_{\pi_l} = N_j$
- N_{π_i} and $N_{\pi_{i+1}}$ are linked by an arc of the tree for $i = 1, ..., l-1$
- $N_{\pi_i} \neq N_{\pi_j}$ for $i \neq j$.

In the following we consider LCMT networked systems. It is worth underlining that a rooted tree is a pair made of a tree and one of its nodes N_r, named as "root". Hence, since a tree is a connected graph, in a LCMT network there is always a path between two nodes and, since there are no cycles, such a path is also unique.

The presence of a special node labeled as "root" induces a natural relation of "order" among the nodes in the following way

Definition 6: Given a rooted tree, consider the path from N_r to another node N_j. A node N_i is said to be an ancestor of N_j if $N_i \neq N_j$ and if it belongs to the path from N_r to N_j. Alternatively, we say that N_j is a descendant of N_i. We also say that N_i is parent of N_j (or that N_j is a child of N_i) if, in addition, N_j and N_i are connected by an arc.

It is straightforward to prove that the root is an ancestor to all the other nodes and that every node but the root has exactly one parent. Hereafter an important result about the correlation property in a LCMT is introduced.

Lemma 7: Given a LCMT $\mathcal{T}$, consider a node N_j and a node $N_i \neq N_j$ which is not a descendant of N_j. Then it holds that $E[\varrho_j X_i] = 0$.

Proof: Let N_r be the root of $\mathcal{T}$ and $N_{\pi_1}, ..., N_{\pi_l}$ the path from N_r to N_i. Exploiting the linear dependencies among the signals of the LCMT, X_i can be expressed in terms of the noises $\varrho_{\pi_1}, ..., \varrho_{\pi_l}$

$$X_i = \sum_{q=1}^{l} W_{i\pi_q}\varrho_{\pi_q} \tag{11}$$

where

$$W_{i\pi_q} = \prod_{h=q}^{l-1} H_{\pi_h}. \tag{12}$$

Since N_i is not a descendant of N_j and $N_i \neq N_j$, we have that $\varrho_{\pi_q} \neq \varrho_j$ for $q = 1, ..., l$, thus

$$E[\varrho_j X_i] = E\left[\varrho_j \sum_{q=1}^{l} W_{i\pi_q}\varrho_{\pi_q}\right] = 0 \tag{13}$$

■

The two following lemmas provide two important inequalities about the coherence functions related to the network signals.

Lemma 8: Consider a LCMT $\mathcal{T}$ and three nodes N_i, N_j and N_k such that

- N_k is a descendant of N_j
- N_i is not a descendant of N_j and $N_i \neq N_j$.

Then we have that $C_{X_iX_j} \geq C_{X_iX_k}$. Moreover, if $\mathcal{T}$ is well-posed then the inequality is strict.

Proof: Consider the path from N_j to N_k described by the sequence $N_{\pi_1}, ..., N_{\pi_l}$. Exploiting the linear relations (1), the process X_k can be expressed in terms of X_j and of the noises acting on the nodes $N_{\pi_2}, ..., N_{\pi_l}$ which are all descendants of N_j.

$$X_k = W_{k\pi_1} X_j + \sum_{q=2}^{l} W_{k\pi_q} \varrho_{\pi_q} \tag{14}$$

where $W_{i\pi_q}$ is defined as in (12). Now, we intend to evaluate the coherence between X_i and X_j. From the assumption on N_i, it follows that N_i is not on the path from N_j to N_k. In other words, N_i is not a descendant of N_{π_q} and $N_i \neq N_{\pi_q}$ for $q = 1, ..., l$. We can write

$$C_{X_iX_k} = \frac{|\Phi_{X_iX_k}|^2}{\Phi_{X_i}\Phi_{X_k}} = \\ = \frac{|W_{k\pi_1}|^2 |\Phi_{X_iX_j}|^2}{\Phi_{X_i}[\Phi_{X_j}|W_{k\pi_1}|^2 + \sum_{q=2}^{l} |W_{k\pi_q}|^2 \Phi_{\varrho_{\pi_q}}]} \tag{15}$$

where the last equality holds because of Lemma 7 . Collecting the factor $\Phi_{X_j}|W_{k\pi_1}|^2$, we obtain

$$C_{X_iX_k} = \frac{|\Phi_{X_iX_j}|^2}{\Phi_{X_i}\Phi_{X_j}\left[1 + \frac{\sum_{q=2}^{l} |W_{k\pi_q}|^2 \Phi_{\varrho_{\pi_q}}}{\Phi_{X_j}|W_{k\pi_1}|^2}\right]} \leq C_{X_iX_j} \tag{16}$$

where the inequality is strict if $\sum_{q=2}^{l} |W_{k\pi_q}|^2 \Phi_{\varrho_{\pi_q}} > 0$. ■

Lemma 9: Consider a LCMT $\mathcal{T}$ and three different nodes N_i, N_j and N_k such that

- N_k is a child of N_j
- $N_i \neq N_j, N_k$ and it is not a descendant of N_k

Then $C_{X_jX_k} \geq C_{X_iX_k}$. Moreover, if $\mathcal{T}$ is well-posed the inequality is strict.

Proof: Assume that $X_k = H_{kj}X_j + \varrho_k$ and let us distinguish two possible scenarios.

case A

First, consider the case where N_j is a descendant of N_i. Consider the path from N_i to N_j described by the sequence of l nodes $N_{\pi_1}, ..., N_{\pi_l}$ where $N_{\pi_1} = N_i$ and $N_{\pi_l} = N_j$. The process X_j can be expressed in terms of X_i and of the noises acting on the nodes $N_{\pi_2}, ..., N_{\pi_l}$ which are all descendants of N_i.

$$X_j = W_{j\pi_1} X_i + \sum_{q=2}^{l} W_{j\pi_q} \varrho_{\pi_q} \,. \tag{17}$$

Exploiting Lemma 7 we can evaluate the following quantities

$$C_{X_iX_k} = \frac{|\Phi_{X_iX_k}|^2}{\Phi_{X_i}\Phi_{X_k}} = \frac{|W_{j\pi_1}|^2 |H_{kj}|^2 |\Phi_{X_i}|^2}{\Phi_{X_i}\Phi_{X_k}} = \\ = \frac{|W_{j\pi_1}|^2 |H_{kj}|^2 \Phi_{X_i}}{\Phi_{X_k}} \tag{18}$$

and

$$C_{X_jX_k} = \frac{|\Phi_{X_jX_k}|^2}{\Phi_{X_j}\Phi_{X_k}} = \frac{|H_{kj}|^2 |\Phi_{X_j}|^2}{\Phi_{X_j}\Phi_{X_k}} = \\ = \frac{|H_{kj}|^2}{\Phi_{X_k}} \left[\Phi_{X_i}|W_{j\pi_1}|^2 + \sum_{q=2}^{l} |W_{j\pi_q}|^2 \Phi_{\varrho_{\pi_q}}\right] \tag{19}$$

By inspection we have the assertion.

Now we are left to consider the case where N_j is not a descendant of N_i. Then, also N_k is not a descendant of N_i. By hypothesis, N_i is not a descendant of N_k, either. Thus, they must have a common ancestor N_d, such that the two paths from N_d to N_k and from N_d to N_i have only N_d in common. Consider the path from N_d to N_i, such that it is possible to write

$$X_i = W_{i\pi_1} X_d + \sum_{q=2}^{l} W_{i\pi_q} \varrho_{\pi_q}. \tag{20}$$

Exploiting lemma 7, we have

$$C_{X_iX_k} = \frac{|\Phi_{X_iX_k}|^2}{\Phi_{X_i}\Phi_{X_k}} = \tag{21}$$

$$= \frac{|\Phi_{X_kX_d}|^2}{\Phi_{X_k}\left[\Phi_{X_d} + \sum_{q=2}^{l} |W_{j\pi_q}|^2 \Phi_{\varrho_{\pi_q}}\right]} \leq \tag{22}$$

$$\leq C_{X_kX_d} \tag{23}$$

If $N_d = N_j$, we have the assertion. If $N_d \neq N_j$, then N_j must be a descendant of N_d. We are in a situation equivalent to case A: there is a node N_d such that N_j is one of its descendants. As a consequence, we can state that

$$C_{X_kX_d} \leq C_{X_kX_j}. \tag{24}$$

Combining the last two inequalities, we conclude that the lemma holds also in this case. ■

All the previous lemmas are functional to the show that the coherence distance (10) is minimal between two contiguous nodes, as summarized in this theorem.

Theorem 10: Given a LCMT $\mathcal{T}$, consider a node N_a and a node $N_b \neq N_a$ which is not directly linked to it. Then there exists a node N_c directly linked to N_a such that

$$d(N_a, N_c) \leq d(N_a, N_b) \tag{25}$$

where the inequality is strict if $\mathcal{T}$ is well-posed.

Proof: First, consider the case where N_b is a descendant of N_a. Name N_c the child of N_a on the path linking it to N_b. Since N_c is directly linked to N_a, we have $N_b \neq N_c$. Moreover N_b is a descendant of N_c. We are allowed to apply lemma (8) with $N_i = N_a$, $N_j = N_c$ and $N_k = N_b$ to have the assertion.

Now, consider the case where N_b is not a descendant of N_a. N_a can not be the root, otherwise N_b would be one of its descendants. Thus N_a has a parent and let us name it N_c. N_b can not be N_c because it is not directly linked to N_a. Applying lemma (9) with $N_i = N_b$, $N_j = N_c$ and $N_k = N_a$ and by the definition of the coherence distance (10), we have the assertion.

■

Theorem 10 can be fruitfully exploited to determine whether two processes in a well-posed LCMT are directly linked. Nonetheless, when we are dealing with data sampled from actual systems the computation of d, that is of the coherence function, is affected by the limited time horizon of the observations. However, the estimates of the spectral and cross-spectral densities converge to the actual values as the time horizon approaches infinity. Hence, in the following we will assume to sample the processes over a sufficiently large time interval.

We are ready to show the main contribution of the paper.

Theorem 11: Consider a well-posed LCMT $\mathcal{T}$ and assume to observe the signals X_j during a time horizon t. Compute an estimate of the coherence based distances $d_{ij} = d(X_i, X_j)$ among the nodes N_j and evaluate the relative Minimum Spanning Tree (MST). When t approaches infinity, the corresponding topology is equivalent to the unique MST T associated to the coherence metric.

Proof: The proof consists in showing that the MST T associated to the distance (10) is unique and corresponds to the LCMT topology. We will prove this result by induction on the number n of nodes of the LCMT.

The basic induction step consists in observing that theorem is true for $n = 2$.

Now assume the theorem true for a LCMT with n nodes. Given a LCMT $\mathcal{T}$ with $n + 1$ nodes, remove one of its "leaves". By leaf we mean a non-root node with no descendants. This operation is always possible since any rooted tree with at least two nodes has at least one leaf. Without loss of generality, let the removed leaf be N_{n+1} and let N_i be its parent. Now we have a LCMT $\mathcal{T}'$ with n nodes and with the same topology of $\mathcal{T}$ apart from the removed arc $(i, n + 1)$. Using the induction hypothesis, we know that the topology of $\mathcal{T}'$ is given by the unique MST T' obtained considering the distances among the nodes $N_1, ..., N_n$. Now compute

$$i^* = \arg \min_{j<N+1} d(X_i, X_{n+1}). \tag{26}$$

The solution of such a minimization problem is unique since the LCMT $\mathcal{T}$ is well posed. Because of lemma 10, the arc $(i^*, N + 1)$ belongs to the topology of $\mathcal{T}$, so we conclude $i^* = i$. Let T be the spanning tree obtained by adding the arc $(i, N + 1)$ to T'. So far, we have shown that T represents the topology of $\mathcal{T}$. We have to prove that T is the unique MST related to the distance (10) among the nodes $N_1, ..., N_{n+1}$. Suppose, by contradiction, that there is a minimum spanning tree $\bar{T} \neq T$ with weight lesser or equal than the weight of T. The only arc of $\bar{T}$ incident to the node N_{n+1} is $(i, n + 1)$. If there were another arc $(k, n + 1)$ in $\bar{T}$ we could replace it with the arc (k, i) obtaining a spanning tree with inferior cost. Indeed, by lemma 9, we would have

$$d(X_k, X_i) < d(X_{n+1}, X_i). \tag{27}$$

So, if $\bar{T}$ is a minimum spanning tree, then X_{n+1} can be connected only to X_i. Let $\bar{T}'$ be the tree obtained by $\bar{T}$ removing the arc $(i, n + 1)$. $\bar{T}'$ is the minimum spanning tree for the nodes $N_1, ..., N_n$ since it has been obtained from $\bar{T}$ removing the node N_{n+1} which has a single connection. However, by the induction hypothesis, there is a unique MST T' among the nodes $N_1, ..., N_n$. Thus we have that $\bar{T}' = T'$. It immediately follows the contradiction that $\bar{T} = T$. ■

So far, we have assumed that the dynamics of the network is described by a rooted tree. Moreover, the previous theorem proves that the topology structure can be correctly identified evaluating the MST according to the distance (10). However, no information is recovered about the root node. The following result shows that such an information is not necessary (or, equivalently, not recoverable). Indeed, from a modeling point of view, the choice of the root can be arbitrary (as long as we are considering non-causal transfer functions linking the processes X_j).

Theorem 12: Given a LCMT $\mathcal{T}$ whose root is the node N_j and given one of its children N_i, it is possible to define another LCMT $\mathcal{T}^*$ with the same tree structure and described by the same processes X_k, $k = 1, ..., n$, such that its root is N_i.

Proof: Consider the Wiener Filter W_{ji} modeling the signal X_j, seen as the output, when X_i is the input

$$X_j = W_{ji} X_i + e_{ji}. \tag{28}$$

Now, consider a rooted tree with the same topology of $\mathcal{T}$ but with N_i as the root. Define $H_k^* = H_k$ and $\varrho_k^* = \varrho_k$ for all $k \neq i, j$. Conversely, define

$$H_j^* = W_{ji} \qquad \varrho_j^* = e_{ji} \tag{29}$$

$$H_i^* \equiv 0 \qquad \varrho_i^* = X_i. \tag{30}$$

To show that the new dynamical network with N_i as root and described by the filters H_k^* is an LCMT, we need to prove that, for $h \neq k$,

$$E[\varrho_h^* \varrho_k^*] = 0. \tag{31}$$

There are three possible scenarios.

If $h = i$ and $k = j$ or $h = j$ and $k = i$, then

$$E[\varrho_h^* \varrho_k^*] = 0. \tag{32}$$

because of the Wiener Filter properties.

If $h = i, j$ and $k \neq i, j$ (or equivalently $h \neq i, j$ and $k = i, j$), then lemma 7 can be applied.

If $h \neq i, j$ and $k \neq i, j$, then

$$E[\varrho_h^* \varrho_k^*] = E[\varrho_h \varrho_k] \tag{33}$$

and we have the assertion because ϱ_h and ϱ_k are two noise signals of the original LCMT $\mathcal{T}$. ■

It is straightforward to show that, starting from an LCMT $\mathcal{T}$, we can arbitrary define a LCMT $\mathcal{T}^*$ having an arbitrary node as root. Indeed, it is sufficient to iteratively apply Theorem 12 along the path starting from the original root to the new one.

IV. Numerical examples

In this section we introduce a suitable framework to illustrate the application of the previous theoretical results to numerical analysis. It is worth observing that the previous results have been developed for the most general class of linear models. Indeed, no assumptions have been done on the order and causality property of the considered transfer functions.

Moreover, let us highlight that the coherence based analysis

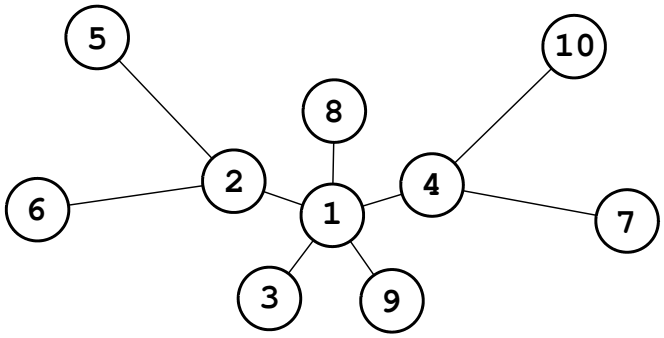

Fig. 1. The figure illustrates the topology of the 10 nodes network analyzed in the numerical examples paragraph. Each node is responsible for a process X_j, while the arcs describe the connections among them, according to the linear SISO model (1). For the data generation we have considered only transfer functions of at most the second order. The noises ϱ_j have been assumed to provide half the power of the affected processes. The samples have been collected over 1000 time steps.

must be realized "off-line", since the processes have to be evaluated over their entire time span. Thus, because the coherence function can be numerically computed only over limited intervals, in the following examples we will consider sufficiently long time spans to reduce the numerical error.

Hence, let us build the original dynamical networks according the following rules:

- each system is described according to the model (1);
- each transfer function H_j is randomly generated and such that it is causal and at most of the second order;
- the tree topology is randomly chosen;
- the noises ϱ_j are numerically generated with a pseudo-random algorithm;
- the noise-to-signal ratio of each system is equal to one.

Then, such networks are simulated over 1000 time steps and the related data X_j are collected. The corresponding coherence based distances are evaluated and used for the extraction of the MST, that defines the link topology.

The above procedure will be first applied to a ten node network. In particular, to test the numerical reliability of the topological identification technique, we repeat such analysis several times, so that a significant number of network configurations is considered. The corresponding results fit the expectations and the real topology is correctly identified each time. In Fig. 1 one of the considered network configurations is depicted, while the related coherence based distance matrix is reported in Table I.

To provide a further test, a new set of similar simulations is performed with a network of fifty dynamical systems, under the same assumptions used in the previous case. Figure 2 presents one of the considered network configurations. For a space limitation issue, we do not report in this manuscript the corresponding coherence based distance matrix. Nonetheless, the computation of the related MST has successfully identified the real network topology in any of the performed simulations.

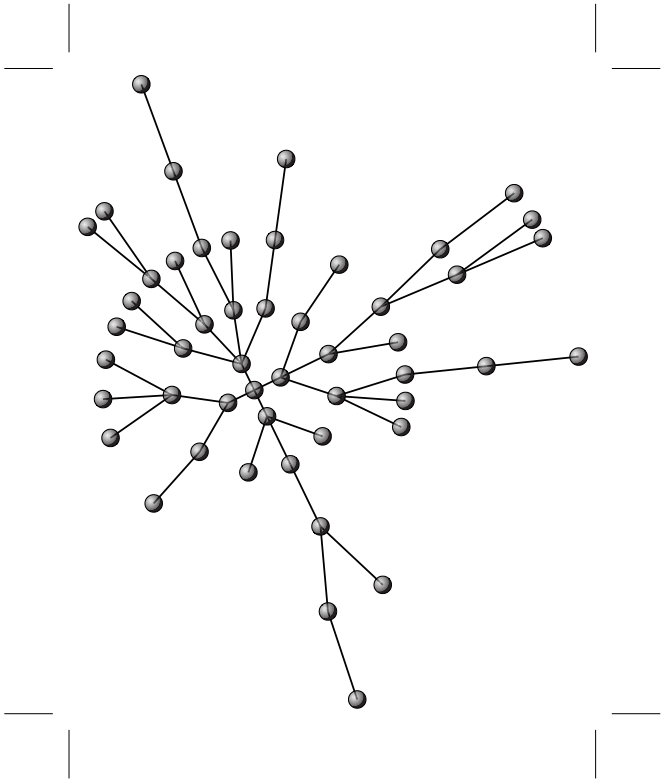

Fig. 2. A representative topological configuration of the 50 nodes network case considered of the numerical examples paragraph. The example has been designed according to the same assumptions of the ten node network of Figure 1.

V. Stock market analysis

In the previous section we have illustrated how the distance (10) can be successfully exploited to derive the exact topology of a tree network of linear systems affected by additive noises. Nonetheless, since the above identification technique is able to catch the most important linear dependencies with respect to the modeling error (4), in the following we present the results obtained by the application of the previous method to the stock market, that is a network of nonlinear systems characterized by multiple dependencies.

Financial systems are, in general, very complex and deriving information from stock markets is a formidable and challenging task, indeed. Moreover, it might seem very reductive the attempt to describe the dependencies among the price trends in terms of linear SISO systems with a tree topology. In fact, we should definitely expect multiple input influences, nonlinear relations and feedbacks. However, we can think of adopting a LCMT in order to detect what are the "strongest" links in the network. As noted in [3], such an information could be usefully exploited to check if a given portfolio is balanced or not. In the following, we report the results obtained by the application of our identification technique.

A collection of 100 stocks of the New York Stock Exchange has been observed for four weeks (twenty market days), in the lapse 03/03/2008 - 03/28/2008 sampling their prices every 2 minutes. The stocks have been chosen on the first 100 stocks with highest trading volume according to the Standard

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
X_1	0	0,7299	0,6675	0,7351	0,8316	0,8542	0,8297	0,7055	0,6549	0,8298
X_2	0,7299	0	0,8065	0,8353	0,6934	0,7358	0,8786	0,8483	0,8299	0,8717
X_3	0,6675	0,8065	0	0,8216	0,8744	0,8807	0,8750	0,8262	0,7841	0,8821
X_4	0,7351	0,8353	0,8216	0	0,8662	0,8722	0,7404	0,8502	0,8198	0,7039
X_5	0,8316	0,6934	0,8744	0,8662	0	0,8540	0,8919	0,8995	0,8730	0,8846
X_6	0,8542	0,7358	0,8807	0,8722	0,8540	0	0,8934	0,8984	0,8796	0,8944
X_7	0,8297	0,8786	0,8750	0,7404	0,8919	0,8934	0	0,8838	0,8694	0,8346
X_8	0,7055	0,8483	0,8262	0,8502	0,8995	0,8984	0,8838	0	0,8167	0,8908
X_9	0,6549	0,8299	0,7841	0,8198	0,8730	0,8796	0,8694	0,8167	0	0,8715
X_{10}	0,8298	0,8717	0,8821	0,7039	0,8846	0,8944	0,8346	0,8908	0,8715	0

TABLE I
THE COHERENCE BASED DISTANCE MATRIX ASSOCIATED TO THE NETWORK TOPOLOGY DEPICTED IN FIG. 1

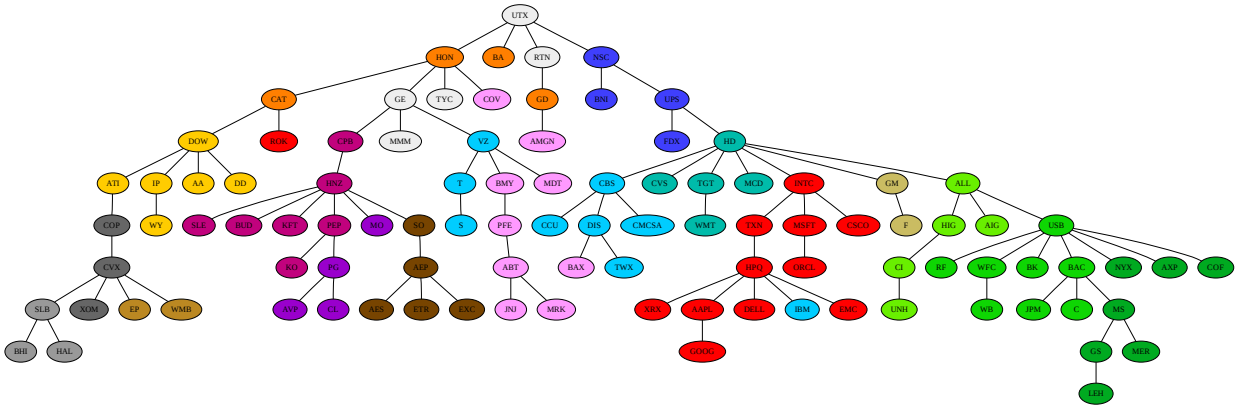

Fig. 3. (color on line) The tree structure obtained using the proposed identification technique. Every node represents a stock and the color represents the business sector it belongs to. The considered sectors are `Basic Material` (yellow), `Conglomerates` (white), `Healthcare` (pink), `Transportations` (dark blue), `Technology` (red), `Capital Goods` (orange), `Utilities` (brown tints), `Consumer` (violet tints), `Financial` (green tints), `Energy` (gray tints) `Services` (light blue tints). Using the industry classification given by Google, the `Financial` sector has also been differentiated among Insurance Companies (light green), Banks (average green) and Investment Companies (dark green); `Services` have been divided in Information Technology (cyan) and Retail (aquamarina), `Consumer` in Food (plum) and Personal-care (purple); `Energy` in Oil & Gas (dark gray) and Well Equipment (light gray); Utilities in Electrical (dark brown) and Natural Gas (light brown).

& Poor Index at the first day of observation and they are reported in Table II. An a-priori organization of the companies has been assumed in accordance with the sector and industry group classification provided by Google Finance®, that is also the source of our data. The whole observation horizon spans almost the whole month of March. Hence, the corresponding price series can not be considered stationary and the statistical tools can not be successfully employed to analyze the raw data. In literature a variety of techniques for the suppression of trends and periodic components in non-stationary time series exists. However, we want to stress that the application of such procedures introduces an additional prefiltering phase, which is responsible for the computational burden increase. Moreover, due to the pre- and post-market sessions, there is a discontinuity between the end value of a day and the opening price of the next one. We have avoided those problems observing that the observation horizon is naturally divided into subperiods, namely weeks and days. In addition, a single market session can be considered a time period sufficiently short to assume that the influence of trends and seasonal factors are negligible. Thus, in our analysis, we have followed the natural approach of dividing the historical series into twenty subperiods corresponding to single days. Then, we considered the sessions separately, i.e. we have computed the coherence-based distances (10) among the stocks for every single day. Finally, we have averaged such daily distances over the whole observation horizon and the related results have been exploited to extract the MST, providing the corresponding market structure.

We find useful to remark that the computation of the distances for smaller data sets is also better performing and that the averaging procedure provides the desired rejection of trends and seasonal components. Notably, a similar idea, even if more sophisticated, is at the basis of the method developed in [21] to detrend non-stationary time series.

The final topology is shown in Figure 3. Every node represents a stock and the color represents the business sector or industry it belongs to. We note that the stocks are very satisfactorily grouped according to their business sectors. We stress that the a-priori classification in sectors is not a hard fact by itself and we are not trying to match it exactly. A company could well be categorized in a sector because of its business, but, at the same time, could show a behaviour similar to and explainable through the dynamics of other sectors. Actually, we would be very interested into finding results of this kind. Indeed, in those very cases, our quantitative analysis would provide the greatest contributions detecting in an objective way something which is “counter-intuitive”. Thus, we just use such a-priori classification as a tool to check if the final topology makes sense and if, at a general level, our approach

provides useful results. Despite this disclaim, it is worth noting that the Financial (green tints), Consumer (violet tints), Basic Materials (yellow), Energy (gray tints) and Transportation (dark blue) sectors are all perfectly grouped, with no exceptions. In Fig. 3, we note a subclusterization of the Financial sector, as well. The Consumer sector shows another prominent subclusterization in the Food (plum) and Personal/Healthcare (purple) industries, while the Energy sector presents an evident subclusterization into the Oil & Gas (dark gray) and Oil Well Equipment (light gray). The Utilities/Electricity companies (dark brown) are, interestingly, a different group. We also observe a big cluster of companies classified as Services (light blue tints). We have differentiated them in the two industries Retail and Information Technology using two slightly different colors, respectively aquamarine and cyan. We also note the presence of three Services companies which are isolated from the other ones: V [Verizon], T [AT&T], and S [Sprint]. All of them are telephone companies. This might suggest that this industry should show at least a slightly different dynamics from the other service companies. Note also how the Technology sector (red) is almost perfectly grouped and how IBM, an IT company, even though classified as a Services company, is located in it. Finally, the two only automobile companies GM and F [Ford] happen to be linked together. The analysis of this four weeks of the month of March cleanly shows a taxonomic arrangement of the stocks even though the choice of a tree structure might have seemed quite reductive at first thought.

VI. Conclusions

This work has illustrated a simple but effective procedure to identify the structure of a network of linear dynamical systems when the topology is described by a tree. To the best knowledge of the authors, the problem of identifying a network has not yet been tackled in scientific literature. The approach followed in this paper is based on the definition of a distance function in order to evaluate if there exists a direct link between two nodes. A few theoretical results are provided, in particular to guarantee the correctness of the identification procedure. An application of the technique to real data has also shown that a tree topology can be sufficient to capture information even in complex situations such as financial stock prices.

Acknowledgments

The authors would like to thank Prof. Tim Sauer for his precious suggestions and advices.

References

[1] R. Rojas, *Neural networks: a systematic introduction*. New York, NY, USA: Springer-Verlag New York, Inc., 1996.

[2] R. Diestel, *Graph Theory*. Berlin, Germany: Springer-Verlag, 2006.

[3] R. N. Mantegna, "Hierarchical structure in financial markets," *Eur. Phys. J. B*, vol. 11, pp. 193–197, 1999.

[4] R. Mantegna and H. Stanley, *An Introduction to Econophysics: Correlations and Complexity in Finance*. Cambridge UK: Cambridge University Press, 2000.

[5] M. Naylora, L. Roseb, and B. Moyle, "Topology of foreign exchange markets using hierarchical structure methods," *Physica A*, vol. 382, pp. 199–208, 2007.

[6] M. Eisen, P. Spellman, P. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proc. Natl. Acad. Sci. USA*, vol. 95, no. 25, pp. 14 863–8, 1998.

[7] E. Ravasz, A. Somera, D. Mongru, Z. Oltvai, and A. Barabï£¡si, "Hierarchical organization of modularity in metabolic networks," *Science*, vol. 297, p. 1551, 2002.

[8] A. Bunn, D. Urban, and T. Keitt, "Landscape connectivity: A conservation application of graph theory," *Journal of Environmental Management*, vol. 59, no. 4, pp. 265 –278, 2000.

[9] P. Monestiez, J.-S. Bailly, P. Lagacheriec, and M. Voltz, "Geostatistical modelling of spatial processes on directed trees: Application to fluvisol extent," *Geoderma*, vol. 128, pp. 179–191, 2005.

[10] D. Urban and T. Keitt, "Landscape connectivity: A graph-theoretic perspective," *Ecology*, vol. 82, no. 5, pp. 1205–1218, 2001.

[11] G. Innocenti and D. Materassi, "A modeling approach to multivariate analysis and clusterization theory," *Journal of Physics A*, vol. 41, no. 20, p. 205101, 2008.

[12] ——, "Topological properties in identification and modeling techniques," *Proc. of the 17th IFAC World Congress, Seoul*, 2008.

[13] J. Banavar, F. Colaiori, A. Flammini, A. Maritan, and A. Rinaldo, "Topology of the fittest transportation network," *Physical Review Letters*, vol. 84, no. 20, pp. 4745–4748, 2000.

[14] J.-S. Bailly, P. Monestiez, and P. Lagacherie, "Modelling spatial variability along drainage networks with geostatistics," *Mathematical Geology*, vol. 38, no. 5, pp. 515–539, 2006.

[15] M. Durand, "Structure of optimal transport networks subject to a global constraint," *Physical Review Letters*, vol. 98, no. 8, p. 088701, 2007.

[16] Z. Shao and H. Zhou, "Optimal transportation network with concave cost functions: loop analysis and algorithms," *Physical Review E*, vol. 75, p. 066112, 2007.

[17] L. Ljung, *System identification: theory for the user (2nd Ed.)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1999.

[18] T. Kailath, A. Sayed, and B. Hassibi, *Linear Estimation*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2000.

[19] C. D. Michener and R. R. Sokal, "A quantitative approach to a problem of classification," *Evolution*, vol. 11, pp. 490–499, 1957.

[20] A. Sayed and T. Kailath, "A survey of spectral factorization methods," *Numerical Linear Algebra with Applications*, vol. 8, pp. 467–469, 2001.

[21] B. Podobnik and H. E. Stanley, "Detrended cross-correlation analysis: A new method for analyzing two non-stationary time series," *Phys. Rev. Lett.*, vol. 100, p. 084102, 2008.

Name	Code	Sector
3M Company	MMM	Conglomerates
Abbott LAboratories	ABT	Healthcare
Aes Corporation	AES	Utilities
Alcoa Inc.	AA	Basic Materials
Allegheny Technologies Inc.	ATI	Basic Materials
Allstate Corporation	ALL	Financial
Altria Group	MO	Consumer/Non-Cyclical
American Electric Power	AEP	Utilities
American Express	AXP	Financial
American International Group	AIG	Financial
Amgen Inc.	AMGN	Healthcare
Anheuser Busch	BUD	Consumer/Non-Cyclical
Apple Inc.	AAPL	Technology
AT&T	T	Services
Avon Products	AVP	Consumer/Non-Cyclical
Baker Hughes Inc.	BHI	Energy
Bank of America	BAC	Financial
Bank of New York Mellon	BK	Financial
Baxter International	BAX	Healthcare
Boeing	BA	Capital Goods
Bristol Myers Squibb	BMY	Healthcare
Burlington Northern Santa Fe	BNI	Transportation
Campbell Soup	CPB	Consumer/Non-Cyclical
Capital One Financial	COF	Financial
Caterpillar Inc.	CAT	Capital Goods
CBS	CBS	Services
Chevron	CVX	Energy
CIGNA	CI	Financial
Cisco Systems	CSCO	Technology
Citigroup Inc	C	Financial
Clear Channel Communications	CCU	Services
Coca-Cola	KO	Consumer/Non-Cyclical
Colgate Palmolive	CL	Consumer/Non-Cyclical
Comcast	CMCSA	Services
Conoco Phillips	COP	Energy
Covidien	COV	Healthcare
CVS Caremark	CVS	Services
Dell Inc	DELL	Technology
Dow Chemical Company	DOW	Basic Materials
E.I. du Pont de Nemours	DD	Basic Materials
El Paso	EP	Utilities
EMC	EMC	Technology
Entergy	ETR	Utilities
Exelon	EXC	Utilities
Exxon Mobil	XOM	Energy
FedEx	FDX	Transportation
Ford Motor	F	Consumer Cyclical
General Dynamics	GD	Capital Goods
General Electric	GE	Conglomerates
General Motors	GM	Consumer Cyclical
Goldman Sachs Group	GS	Financial
Google Inc.	GOOG	Technology
Halliburton	HAL	Energy
Hartford Financial Services	HIG	Financial
H. J. Heinz	HNZ	Consumer/Non-Cyclical
Hewlett-Packard	HPQ	Technology
Home Depot	HD	Services
Honeywell International	HON	Capital Goods
Intel	INTC	Technology
International Business Machines	IBM	Services
International Paper	IP	Basic Materials
Johnson & Johnson	JNJ	Healthcare
JPMorgan Chase	JPM	Financial
Kraft Foods	KFT	Consumer/Non-Cyclical
Lehman Brothers Holding	LEH	Financial
McDonald's	MCD	Services
Medtronic	MDT	Healthcare
Merck	MRK	Healthcare
Merril Lynch	MER	Financial
Microsoft	MSFT	Technology
Morgan Stanley	MS	Financial
Norfolk Souther Group	NSC	Transportation
NYSE Euronext	NYX	Financial
Oracle	ORCL	Technology
Pespi	PEP	Consumer/Non-Cyclical
Pfizer Inc.	PFE	Healthcare
Procter & Gamble	PG	Consumer/Non-Cyclical
Raytheon	RTN	Conglomerates
Regions Financial	RF	Financial
Rockwell Automation	ROK	Technology
Sara Lee	SLE	Consumer/Non-Cyclical
Schlumberger Limited	SLB	Energy
Southern	SO	Utilities
Sprint Nextel	S	Services
Target	TGT	Services
Texas Instruments Inc.	TXN	Technology
Time Warner	TWX	Services
Tyco International	TYC	Conglomerates
U. S. Bancorp	USB	Financial
United Parcel Service	UPS	Transportation
United Technologies	UTX	Conglomerates
UnitedHealth Group Inc.	UNH	Financial
Verizon Communications	VZ	Services
Wachovia	WB	Financial
Wal-Mart Stores	WMT	Services
Walt Disney	DIS	Services
Wells Fargo	WFC	Financial
Weyerhaeuser Company	WY	Basic Materials
Williams Companies	WMB	Utilities
Xerox	XRX	Technology

TABLE II
LIST OF THE COMPANIES CONSIDERED IN THE ANALYSIS